2010 Digital Universe Study
iView content – FINAL
Version: 4-26-2010

Title: A Digital Universe Decade – Are You Ready?

Tab 1: The Digital Universe Decade

"You Ain't Seen Nothing Yet." The title of that track from the 1974 Bachman-Turner Overdrive album *Not Fragile* aptly describes the state of today's Digital Universe. Between now and 2020, the amount of digital information created and replicated in the world will grow to an almost inconceivable 35 trillion gigabytes as all major forms of media – voice, TV, radio, print – complete the journey from analog to digital.

At the same time, the influx of consumer technologies into the workplace will create stresses and strains on the organizations that must manage, store, protect, and dispose of all this electronic content. So, if you have ever suffered from information overload or been bombarded with emails, texts, instant messages, documents, pictures, videos, and social network invitations, get ready, this is just the beginning.
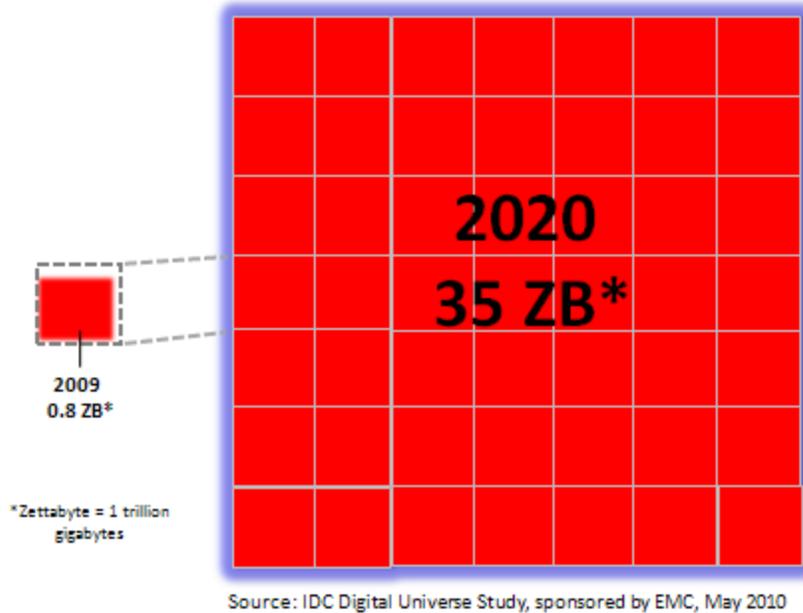
Since 2007, on behalf of EMC Corporation, IDC has been sizing what it calls the Digital Universe, or the amount of digital information created and replicated in a year.

Here are just a few points to whet your appetite for the rest of the tabs in this IDC iView:

- Last year, despite the global recession, the Digital Universe set a record. It grew by 62% to nearly 800,000 petabytes. A petabyte is a million gigabytes. Picture a stack of DVDs reaching from the earth to the moon and back.
- This year, the Digital Universe will grow almost as fast to 1.2 *million* petabytes, or 1.2 zettabytes. (There's a word we haven't had to use until now.)
- This explosive growth means that by 2020, our Digital Universe will be 44 TIMES AS BIG as it was in 2009 (Figure 1). Our stack of DVDs would now reach halfway to Mars.

Figure 1: The Digital Universe 2009 – 2020
Growing by a Factor of 44

2020

35 ZB*

2009
0.8 ZB*

*Zettabyte = 1 trillion gigabytes

Source: IDC Digital Universe Study, sponsored by EMC, May 2010

Here's another question for you. What comes after a quadrillion? That's right, a quintillion, an incomprehensible number, yet the one you need to describe the number of information containers – packets, files, images, records, signals – that the bits in the Digital Universe will be in by 2020. There will be 25 quintillion containers.

These containers, the files if you will, are the things that are actually managed, protected, and stored in the Digital Universe.

And, because of the growth of embedded systems in the smart grid, smart cities, logistic item tracking, and so on, the average file size in the universe is getting smaller. The number of things to be managed is growing twice as fast as the total number of gigabytes. Good luck, all you CIOs out there.

Think of the growth of the Digital Universe as a perpetual tsunami. As this universe grows by an order of magnitude, we will have to deal with information in new ways:

❑ How will we find the information we need when we need it? We will need new search and discovery tools. Most of the Digital Universe is unstructured data (for example, images and voice packets). We will need new ways to add structure to unstructured data, to look INSIDE the information containers and recognize content such as a face in a security video. In fact, the fastest-growing category in the Digital Universe is metadata, or data about data.

❑ How will we know what information we need to keep, and how will we keep it? Yes, we will need new technical solutions tied to storage, but we will surely also need new ways to *manage* our information. We'll need to classify it by importance, know when to delete it, and predict which information we will need in a hurry.

❑ How will we follow the growing number of government and industry rules about retaining records, tracking transactions, and ensuring information privacy? Compliance

with regulations has become an entire industry – a $46 billion industry last year – but will it be enough?

❑ How will we protect the information we need to protect? If the amount of information in the Digital Universe is growing at 50% a year or so, the subset of information that needs to be secured is growing almost twice as fast. The amount of UNPROTECTED yet sensitive data is growing even faster.

As we contemplate the growth of the Digital Universe, these are some of the things we need to think about:

➢ New search tools
➢ Ways to add structure to unstructured data
➢ New storage and information management techniques
➢ More compliance tools
➢ Better security

There are plenty of others, including the role of cloud computing, the consumerization of the workplace, the growing share of the Digital Universe coming from China and India, and the growing diversity – in content type and container type – of the Digital Universe.

Here is another statistic to keep in mind before you start reviewing our other findings: Although the amount of information in the Digital Universe will grow by a factor of 44, and the number of containers or files will grow by a factor of 67 from 2009 to 2020, the number of IT professionals in the world will grow only by a factor of 1.4.

Big changes are coming.

Tab 2. Information in the Clouds

By 2020, a significant portion of the Digital Universe will be centrally hosted, managed, or stored in public or private repositories that today we call "cloud services." And even if a byte in the Digital Universe does not "live in the cloud" permanently, it will, in all likelihood, pass through the cloud at some point in its life.

There are almost as many definitions of cloud services as there are vendors trying to gain advantage by offering them. But in the IDC definition, they require availability over a network, consumption on-demand with pay-as-you-go billing, and some level of user control and system openness that separates cloud services from simple online delivery of content. It's software as a service, not downloading software programs. It's watching on-demand Internet TV, not merely downloading Netflix videos.

At the same time, cloud services can be offered as a shared common functionality (public cloud) or as a private version (private cloud), where an organization maintains complete control of all of the IT resources and how they are managed and secured. The latter can even be offered within the enterprise itself.

Although IDC has only sized the market for IT functionality – hardware, software, services – delivered over the public cloud at this point, a look at the makeup of the Digital Universe by content type – entertainment, financial transaction, medical information, and user-generated image content – gives one a feel for how big cloud services in other industries could become in the Digital Universe.

Using reasonable forecast assumptions such as those in the IDC forecast for IT cloud services, it is possible to conclude that as much as 15% of the information in the Digital Universe in 2020 could be part of a cloud service – created in the cloud, delivered to the cloud, stored and manipulated in the cloud, etc. Even more information could "pass through the cloud," that is, be transported using a cloud services email system or shared community, be stored temporarily on disk drives in the cloud, be secured via a cloud service, etc. By 2020, more than a third of all the information in the Digital Universe will either live in or pass through the cloud (Figure 2).

Because so much of the content in the Digital Universe is in the form of images – from digital TV to user-generated images – nearly 50% of the Digital Universe cloud-based subset will be tied to entertainment.

Figure 2: The Digital Universe in the Clouds, 2020
Potential Available Share

Total Digital Universe

Touched By Cloud

Cloud Services
5 ZB

12 ZB

35 ZB
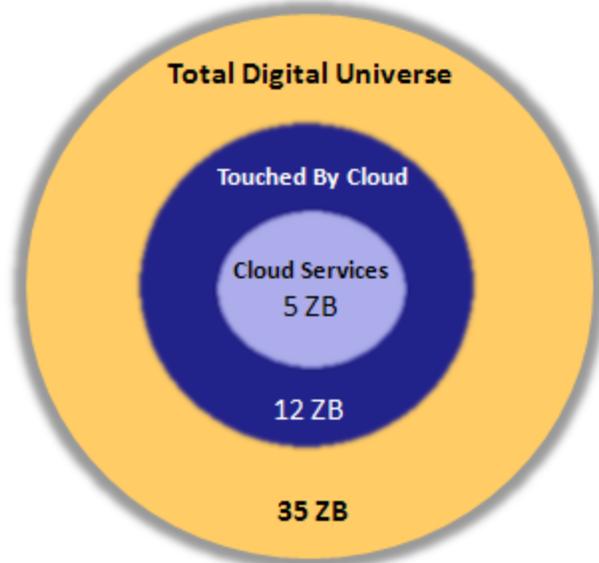
Source: IDC Digital Universe Study, sponsored by EMC, May 2010

But every industry will have some form of cloud-based services to offer. And with these services come unique challenges and benefits:

❑ The promise of cloud services, besides lower upfront investment, is ease of management. But as users of IM systems in the early days and users of iTunes today realize, conversion from one service to another is not always seamless.

❑ Cloud services imply access to broadband services, which are not always available. Perhaps more challenging is the fact that cloud services may spur *more* need for bandwidth. Already, telco carriers are complaining that users of smart phones are using much more bandwidth than predicted as they rush to use more and more applications.

❑ Overall, security may actually be enhanced in the cloud – as cloud providers build security and transparency into the cloud infrastructure from the get-go. But the impact of a failure, should there be one, could be significant. This could be a problem in private clouds as well as public clouds. And with the loss of physical control over infrastructure, ensuring visibility in the cloud will be critical for demonstrating compliance.

Even with information stored or manipulated in the cloud, enterprises still have responsibility and liability over it. Managing this from afar might be a challenge.

But the benefits of cloud services will be substantial. Using IDC's current forecast for IT cloud services and assuming that the use of cloud services could lower the portion of the IT budget devoted to legacy system maintenance by a fraction of a percent, we estimate that the increase in IT dollars spent on innovation could drive more than $1 trillion in increased business revenues between now and the end of 2014.

Tab 3. Protected and Unprotected Data

Do you know where your social security number is?

Think about it. You probably have to enter it onto forms 10 times a year, maybe 50. From there it enters the Digital Universe, living in all sorts of databases, from those in your doctor's office and tax accountant's office, to bank records, company personnel records, mortgage records, and so on. It shuttles around from database to database at the speed of light. It gets backed up again and again.

In fact, those 10 entries (counting the entries into linked or associated files, accessed by people allowed to access the databases with your number in it, and the backups) could be propagated as many as a million times in a year!

How secure are those entries?

By 2020, almost 50% of the information in the Digital Universe will require a level of IT-based security beyond a baseline level of virus protection and physical protection. That's up from about 30% this year. And while the portion of that part of the Digital Universe that needs the *highest* level of security is small – in gigabytes and total files – that portion will grow by a factor of 100.

Not all data needs to be protected equally. A YouTube video of a cat doing tricks would seem to need less protection against hacking or corruption than a home-banking customer's account balances. But each YouTube video is associated with an IP address and end-user profile, and of course, that video might not be of a cat but of something not fit for public viewing. Even worse, that seemingly innocuous cat video may actually be an effective delivery mechanism for the new variants of malware being created by the criminal underground.

For the sake of understanding the degree of security in the Digital Universe, we have classified information that requires security into five successively higher security-level categories:

- ❑ Privacy only – such as an email address on a YouTube upload
- ❑ Compliance-driven – such as emails that might be discoverable in litigation or be subject to retention rules
- ❑ Custodial data – account information, a breach of which could lead to or aid in identity theft
- ❑ Confidential data – information the originator wants to protect, such as trade secrets, customer lists, or confidential memos
- ❑ Lockdown data – information requiring the highest security, such as financial transactions, personnel files, medical records, or military intelligence

Obviously, information can switch categories over its life, or, in aggregation, gain more value over time and hence need higher security. The information that you visited a single website might be less sensitive than your entire web-browsing history, or even less sensitive than the information about how many *times* you visited the website.

By examining information by category and source, it's possible to estimate the amount of information in the Digital Universe that needs some level of security. That which doesn't is

mostly transient data, especially digital TV signals and voice packets that aren't needed after a broadcast or call is over (Figure 3).

Figure 3: The Need for Information Security
Percentage of the Digital Universe
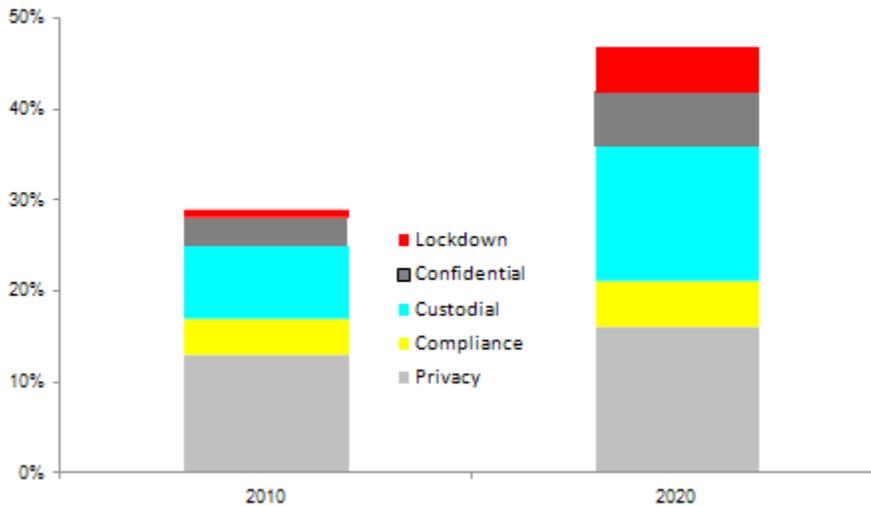
Source: IDC Digital Universe Study, sponsored by EMC, May 2010

However, just because the information *should* have protection doesn't mean that it does. Using the same process of categorization, it is possible to estimate the amount of information that is not adequately protected.

And that amount is growing.

If you look at the information in the Digital Universe that needs to be protected by number of containers or "files" (rather than by number of bytes), the percentage needing protection is more than 90%. And the amount of unprotected data will grow by a factor of 90 between now and 2020 (Figure 4).

The issues for CIOs are pretty clear. They need access to tools and expertise to protect this burgeoning quantity of data needing protection in both the physical and, increasingly, the virtual worlds. But they also need the support of the business units in dealing with the policy and training issues involved – and getting that support has been a perennial problem. Without another regulatory driver such as Sarbanes-Oxley or a catastrophic breach, it is often difficult for CIOs to get the full attention of management regarding the non-technical aspects of information security.

In fact, the issue is even more complex. What a consumer or company wants protected (emails they wish to recall, buying or searching patterns, old Facebook photos) may change from day to day depending on circumstances or because of changes in the originator's own status. (For example, Sarah Palin's email account in the state of Alaska got a lot more interesting to hackers the day she was announced as a Vice Presidential candidate, and it was hacked soon thereafter.)

Figure 4: Unprotected Data Needing Protection
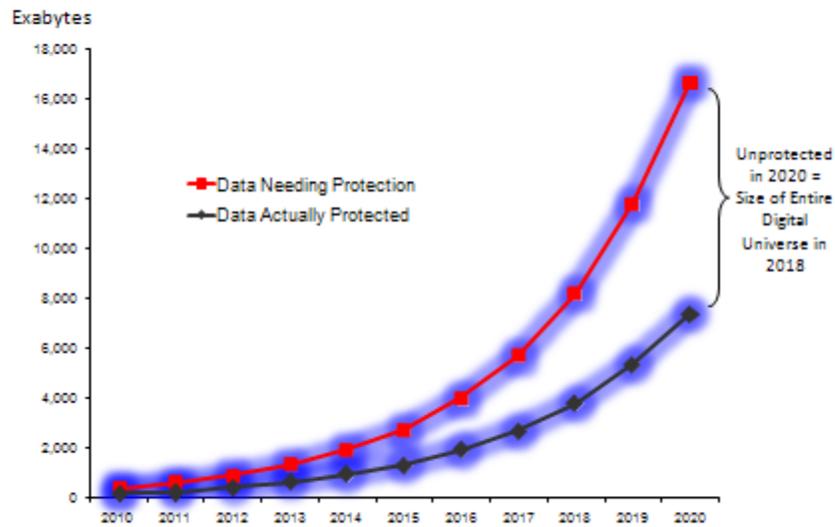
Source: IDC Digital Universe Study, sponsored by EMC, May 2010

Finally, they may not even know there is data about them in the Digital Universe. If they did know about it, they would want it to be protected or otherwise proscribed. In creating the model for the Digital Universe, IDC has discovered that the gigabytes a person may create through his or her own actions – taking photos, blogging, sending emails, getting cash from an ATM, downloading MP3s – is less than 10% the information ABOUT that person in the Digital Universe. The other 90% is composed of credit records, surveillance photos, analytics on behavior, web-use histories, and so on.

Probably EVERY byte in the Digital Universe could use some security and privacy protection. But we will never know because we can never know exactly what all those files and gigabytes actually contain.

Tab 4: The Future of Digital Information Storage

In the inaugural Digital Universe study, IDC forecasted that by 2007, for the first time, the amount of digital information created would exceed the amount of available storage. That inflection point has since become a gulf that continues to expand. Every year, the industry ships thousands of petabytes of new storage capacity, including hard disk drives, optical, tape, nonvolatile memory (flash), and volatile memory (DRAM). The total amount of storage capacity available is equal to all new shipments of storage plus all unused storage from previous years. While previously consumed storage could be overwritten, IDC assumes this capacity is reserved for what is already stored on the storage medium.

Hence, we have a growing gap between the amount of digital content being created and the amount of available digital storage. IDC estimates that in 2009, if people had wanted to store every gigabyte of digital content created, they would have had a shortfall of around 35%. This gap is expected to grow to more than 60% (that is, more than 60% of the petabytes created could not be stored) over the next several years (Figure 5).

Figure 5: The Emerging Gap
Information Creation > Storage Available

Source: IDC Digital Universe Study, sponsored by EMC, May 2010
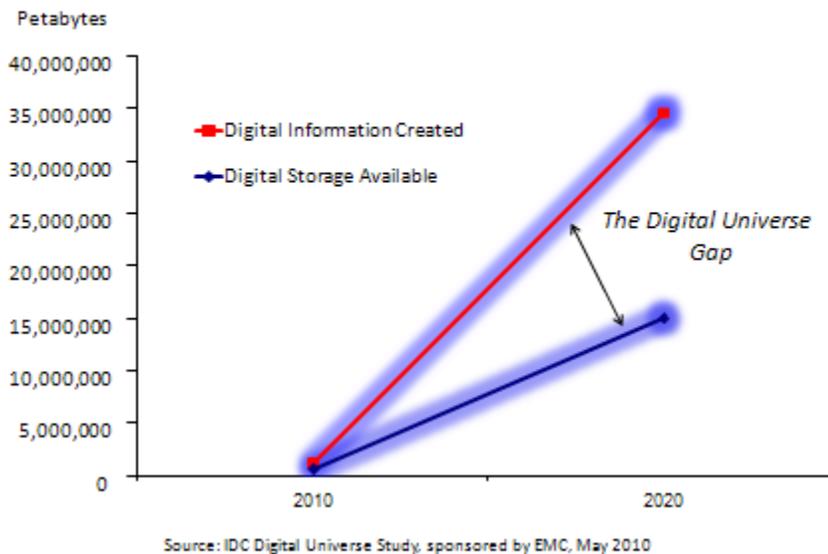
While much of the digital content we create is simply not that important (not much different from the paper magazines and newspapers that we throw away, or the telephone conversations, receipts, bad pictures, etc., that we never save), the amount of data that does require permanent or longer-term preservation for a multitude of reasons is increasing exponentially.

But as we peer into the future, we see the greatest challenges are related not to *how* to store the information we want to keep, but rather to:

- Reducing the cost to store all of this content
- Reducing the risk (and even greater cost) of losing all of this content
- Extracting all of the value out of the content that we save

IDC data shows that nearly 75% of our digital world is a copy – in other words, only 25% is unique. Granted, various laws and regulations require multiple copies exist in order to ensure the availability of data over a long period of time. Multiple copies of data are also necessary to ensure proper performance by applications, or to protect data in the event of hardware failures. Nevertheless, the amount of data redundancy is excessive in many cases, and it represents a prime area for improvement and cost reduction.

Most of today's de-duplication happens on 2nd tier storage, but it doesn't have to. Tomorrow's opportunity is for de-duplication on primary storage (assuming no impact to performance), which would significantly reduce or eliminate post-process de-duplication. The cloud is an especially attractive place to eliminate redundancy, given its one-to-many model of content aggregation.

IDC's forecast portends a huge amount of consumer-attached storage, either directly to hosts or via a network. The external drive market will continue to experience considerable growth. Hence, there is a growing need to enable storage management by consumers. As computing races down the mobility path, one person may find himself with multiple computing devices (each potentially with its own operating system), and various amounts of local storage, all of them providing similar ways to access the cloud (via wireless cellular or wireless broadband).

Sharing the same content among these devices will increase in importance. The cloud must play a strategic role in becoming the central axis for all of this content – yet there is much work to do in order to make this happen. Until then, individuals will strive to keep passwords, content, and services mapped appropriately among their digital device portfolio.

Personalized services (e.g., GPS, proactive coupon pushing, e-commerce) will grow in importance as more data about one's personal habits and preferences is captured and mined by sophisticated applications. Granted, many people today may find this to be intrusive or a direct violation of their personal privacy. But as newer generations of teens and adults use their beloved digital devices, various types of data will be captured and leveraged to deliver services that not only will be embraced, but also will be considered one of those can't-get-along-without-it technologies. The personal data captured and used to deliver these services must be tied to an individual anonymously, and it must be managed in accordance with strict governance and compliance procedures in order to ensure that privacy is not breached. This is, as they say, "easier said than done." Nevertheless, it must be done.

Personal privacy is paramount, and although there are behaviors one can employ to lessen one's digital footprint, no one can completely disappear off the ubiquitous digital grid. Video surveillance and purchasing transactions will be nearly impossible to avoid, and interaction in social networks will be either commonplace or a temptation that is difficult to overcome.

In the end, corporations and consumers will continue to create, copy, and store information – mostly on traditional storage technologies such as hard disk drives. Much of the data that is created will be disposable or a means to a final copy – hence the growing gap between content creation and available storage. However, the content that is stored will be deemed important and vital for evolving our businesses and business models, for extending our digital presence (that should result in more convenience and personalization), and perhaps most important, for protecting our digital heirlooms.

Tab 5: Consumers Without Borders

When you think of the Digital Universe, you may think of the financial databases of Wall Street, the acres of servers operating at giant Internet service providers, or the storage devices supporting 100 million enterprises in the world.
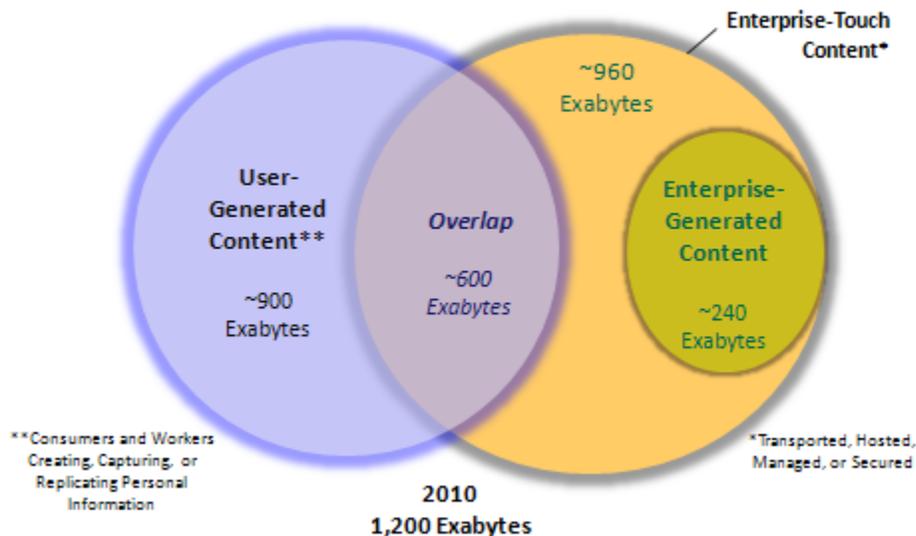
But, in fact, most of the Digital Universe begins with an action by a consumer – an email typed on a laptop, a digital photo taken at a wedding, a movie downloaded from Netflix.

In fact, more than 70% of the Digital Universe this year will be generated by users – individuals at home, at work, and on the go. That's 880 billion gigabytes.

At the same time, most of the gigabytes in the Digital Universe pass through the servers, network, or routers of an enterprise at some point. When they do, the enterprise is responsible at that moment for managing that content, protecting user privacy, watching over account information, and protecting copyright. It was the breach of personal email accounts in China that drove Google off the mainland this year – an excellent example of enterprise liability for consumer-created data.

We classify user-generated content for which enterprises are responsible as "enterprise touch." About two thirds of all user-generated content falls into this category. Here's another way to think of it: While enterprise-generated content accounts for 20% of the Digital Universe, enterprises are liable for 80% (Figure 6).

Figure 6: User Creation = Enterprise Worries

Enterprise-Touch Content*

~960 Exabytes

User-Generated Content**

Overlap

~600 Exabytes

Enterprise-Generated Content

~900 Exabytes

~240 Exabytes

**Consumers and Workers Creating, Capturing, or Replicating Personal Information

*Transported, Hosted, Managed, or Secured

2010
1,200 Exabytes

Source: IDC Digital Universe Study, sponsored by EMC, May 2010

This enterprise liability will only get worse as social networking and Web 2.0 technologies continue to permeate the enterprise. Research by IDC shows that workers who use those technologies at home frequently also use them at work, often commingling their personal and

business content. They use the same smart phones and laptops for work and personal activities. If they blog at home, they want to blog at work.

Many companies are struggling to keep up with this blurring of personal and enterprise boundaries. Only half of the companies surveyed in a separate study by IDC have any kind of corporate guidelines for employee use of social media at work. And nearly two thirds of the enterprise social media activity taking place is being driven by employee initiatives (Figure 7).

This blurring of boundaries is also blending the risks to information. As workers increasingly bring their personal devices and online habits to work, they are also bringing behind the firewall and into the corporate data center the viruses, Trojans, and other malware typically associated with consumer online fraud. This exposes the corporation to increased risk of information theft.

Figure 7: Social Networking Invades the Enterprise

Q. Do employees at your company use social networking for business through a corporate or self-directed initiative?
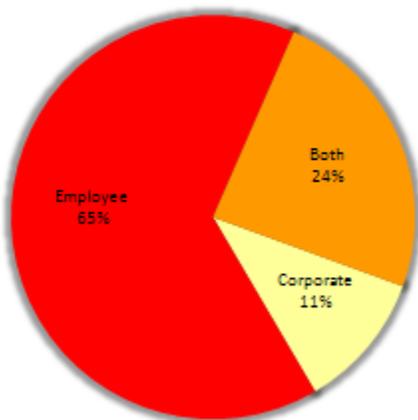
Q. Has your company published official social media guidelines for employees?

Both
24%

Employee
65%

Corporate
11%

No
48%

Don't
Know
10%

Yes
42%
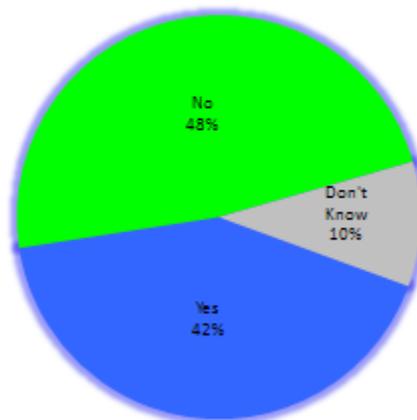
Source: IDC, 2009   7

Employers experience several dilemmas in harnessing social media for the enterprise:

❑ Setting policy. Do you support all employee-created content and devices? Do you let any employee or department set up customer or supplier social media communities? If you set standards, how do you enforce them? Who sets the policy – individual departments, corporate headquarters, IT, business units?
❑ Assuming you see the need for enterprise social networking, how do you make it easy for individuals and departments to engage in Web 2.0 activities – do you offer 100% support or just point them in the right direction?
❑ How do you ensure employee or department social media information is backed up and archived properly, that it adheres to relevant standards and laws, and that it supports the enterprise brand?
❑ How do you ensure security, privacy, and intellectual property protection for these employee-driven activities?

These are all issues now being faced by enterprises. And the social media invasion of the enterprise has just begun. IDC estimates that by 2020, business transactions on the Internet – business-to-business and business-to-consumer – will reach 450 billion a day.

The interactions taking place in the Digital Universe will not only add to its size and growth, but also to the increasing challenge of managing and securing it.

Tab 6. Bucks and Bytes

In 2009, the world spent nearly $4 trillion on hardware, software, services, networks, and IT staff to manage the Digital Universe. That spending is expected to grow modestly between now and 2020, which means the cost of managing each byte in the Digital Universe will drop steadily – an incentive to create even *more* information (Figure 8).

Figure 8: The Decreasing Cost of Managing Information
Total Investment in the Digital Universe

Investment per Gigabyte

$4 Trillion Worldwide in 2009

— Investment per Gigabyte
— Exabytes

Exabytes
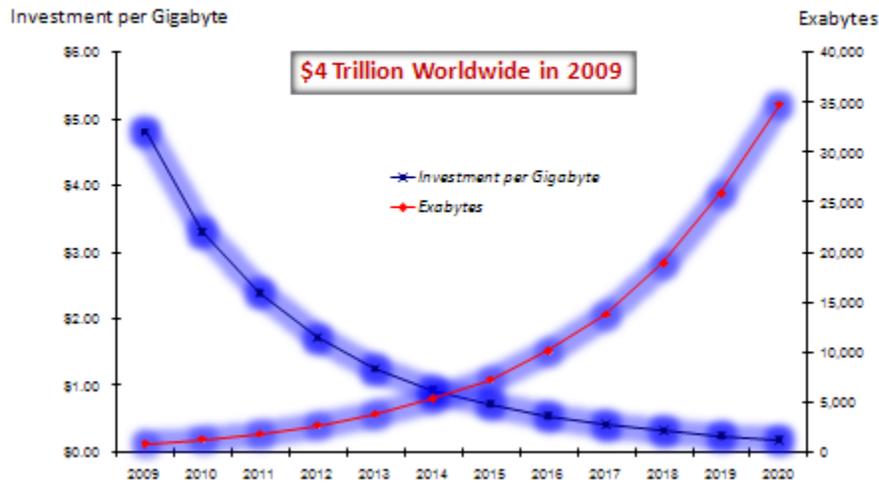
Source: IDC Digital Universe Study, sponsored by EMC, May 2010

There are implications here. While the cost per byte drops, so does the investment in IT staff per byte. The investment in IT staff per information container, or file, drops even faster (Figure 9).

Figure 9: The Decreasing Cost of Managing Information
IT Staff Investment

Source: IDC Digital Universe Study, sponsored by EMC, May 2010
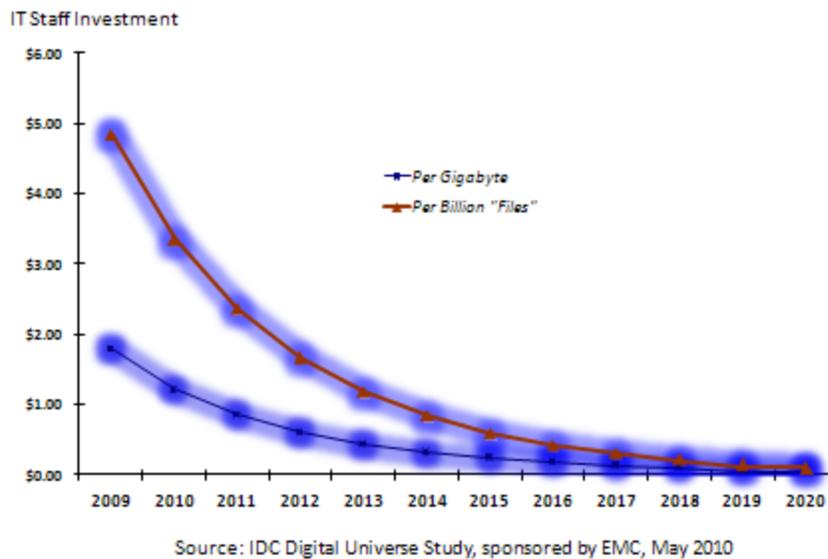
This falling investment ratio, while spurring the growth of the Digital Universe, also means that the tools for managing it will have to change. For instance:

- The increased complexity of managing digital information will be an incentive to migrate to cloud services.
- Within data centers, expect continued pressure for data center automation, consolidation, and virtualization.
- For the management of end-user and customer business transactions, look for more end-user self-service.
- Expect bottlenecks in key specialties such as security (*especially* security), information management, advanced content management, and real-time processing.
- This manage-more-with-less situation will put ever-increasing stress on IT organizations. Those that manage this stress better than competitors will have an advantage.

Tab 7: Call to Action

Between 2009 and 2020, the information in the Digital Universe will grow by a factor of 44; the number of "files" in it to be managed will grow by a factor of 67, and storage capacity will grow by a factor of 30.

Yet the staffing and investment to manage the Digital Universe will grow by a factor of 1.4.

This should make it clear to CIOs and business executives that much of the next 10 years of their careers will be spent dealing with the challenge of the mismatch of these growth rates.

From reviewing the various tabs in this iView, we can narrow the CIO's issues down to:

- Developing tools for search and discovery of information as the Digital Universe expands, including finding ways to add structure to unstructured data though metadata, automatic content tagging, and pattern recognition.
- Deploying tools for new levels of information management and prioritized storage.
- Deploying tools and expertise for security and privacy protection for a growing portion of the Digital Universe in hybrid physical/virtual environments.
- Getting ready for some level of conversion to cloud-based services to start equipping IT staff with the new skills required for providing IT as a service and to obtain some economy of scale for ever-scarcer IT talent.
- Obtaining support from top management and from business units to implement the non-technical aspects of dealing with the Digital Universe, including setting policies on social media, training end-users on information security, and classifying information in an effort to set storage priorities.

As ever, the first order of business is to educate management about the issues and infuse a sense of urgency across the enterprise. As we saw during the recession of 2009, the Digital Universe is a force unto itself. It will grow whether enterprises are ready for it or not.

The second order of business is to plan more than one step out. For instance, if a major new storage management application takes two years to implement – from planning to infrastructure upgrade completion – the Digital Universe will be twice as big at the end of the project as it was at the beginning.

The third order of business is to consider issues of privacy, security, and protection early on and to embed solutions into developing infrastructures, particularly in the virtual space, where we have the opportunity to build-in security from the get-go.

The fourth order of business is to continue working on the relationship with the business units, where most of the funding for new projects originates and where most of the corporate desire to drive policy and training resides.

CIO Action Items

- ➢ **Deploy new IT tools for information management and security.**
- ➢ **Develop a sense of urgency in top management and the business units.**
- ➢ **Develop a long-term plan.**

- ➢ **Increase the bond with the business units.**
- ➢ **Offload work to the cloud.**

In many ways, consumers and employees have some of the same action items when it comes to dealing with their contributions to the Digital Universe. Chief among these is a sense of urgency when it comes to security. Also important is gaining an awareness of how much information *about them* sits in the Digital Universe and how fast that information is growing.

Methodology

Our basic approach of sizing the expanding digital universe was to:

1. Develop a forecast for the installed base of devices or applications that could capture or create digital information.
2. Estimate how many units of information – files, images, songs, minutes of video, phone calls, packets of information – were created in a year.
3. Convert these units to megabytes using assumptions about resolutions, digital conversion rates, and usage patterns.
4. Estimate the number of times a unit of information is replicated, either to share or store. The latter can be a small number, for example, the number of spreadsheets shared, or a large number, such as the number of movies put onto DVD or songs uploaded onto a peer-to-peer network.

Much of this information is part of IDC's ongoing research. For more on the methodology and key assumptions, see the first IDC Digital Universe paper, published in 2007.
http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf

Since 2007, we have added to our forecast about 14 new classes of devices and applications, such as auto/air/marine intelligent systems, building automation, household appliances, digital signage, smart meters, smart thermostats, additional medical electronics and monitors, test and measurement devices, casino/pachinko systems, SMS, information publishers, scientific computing, and metadata